Continuous vs. categorical variations between speakers

Michael Jessen
Department of Speaker Identification and Audio Analysis, Bundeskriminalamt, Germany
Michael.Jessen@bka.bund.de

When average fundamental frequency is investigated based on a large number of speakers and the results are plotted in the form of a histogram, they can be modeled as a single (mono-modal) Gaussian distribution (e.g. Jessen et al., 2005 for German). With this pattern, which will be referred to as continuous speaker variation, speakers differ from each other continuously from low to high values, and most speakers are located somewhere towards the center of this distribution. Since this is a pattern that is found in many biological variables, one explanation is based on the anatomical differences between individuals in terms of (mainly) vocal fold length. However, single Gaussian distributions of speaker differences are also found with f0-variability and with articulation rate (Jessen et al., 2005; Jessen, 2007), which are better explained as habitual rather than anatomically motivated parameters. One question arising from this state of the art is whether all interspeaker variations have this single Gaussian patterning. One domain in which one might expect deviations from this pattern are variations between speakers that are more strongly intertwined with the linguistic system of a language than the parameters mentioned so far. This possibility relates to the question of whether idiolects exist (cf. Nolan, 1994): is it possible that speakers who share the same language variety (e.g. dialect) choose different phonetic categories for the same across-speaker phonological target? If that is the case, one would expect bimodal or multimodal distributions in interspeaker variation, where each peak corresponds to a different phonetic category.

This question was investigated acoustically with two topics in the phonetics and phonology of German. The first topic is referred to in the literature as rounding assimilation (Kloeke, 1982): it is about the issue whether the offglide of the diphthong that occurs in words like Beute ‘prey’ or Bäume ‘trees’ is rounded like its peak (a lax O-target) or unrounded. The second topic is stop epenthesis in words like Zins ‘interest’ or Hals ‘neck’, in which a stop might emerge between the nasal or lateral and the following fricative. Stop epenthesis of this type has been reported for many languages including English (Fourakis and Port, 1986). Pronouncing dictionaries, phonetic textbooks and phonological analyses of German do not agree about the presence vs. absence of rounding assimilation and stop epenthesis. It is possible that this lack of agreement reflects idiolectal variation. Speaker variations of these two phenomena were investigated on the basis of read-sentence material from 100 male speakers of German. Results show that acoustic correlates of rounding assimilation distribute in the familiar mono-modal fashion. Based on this pattern it is argued that speakers aim for the unrounded diphthong offglide and that rounding is introduced gradually in the form of coarticulation rather than assimilation. With stop epenthesis, on the other hand, there was a deviation from the familiar pattern: with the sequence “ns” there was a narrow peak at closure durations close to zero and another, much broader peak for higher values. This was less clear, however, for the sequence “ls”, where values close to zero were less frequent. The patterns of stop epenthesis will be shown in detail and the implications for the questions posed above will be discussed.

References
Realisation of German final syllable /-ən/ as a cue to accent authenticity for French accents in German

Sara Neuhauser and Adrian P. Simpson
Institut für Germanistische Sprachwissenschaft, Universität Jena, Germany
{sara.neuhauser|adrian.simpson}@uni-jena.de

Earlier studies concerning foreign accent imitation (Neuhauser & Simpson, 2007a&b) showed that native speakers of German varied widely in their ability to produce a foreign accent judged by listeners to be authentic. There was also wide variation in the ability of listeners to successfully identify the genuine non-native speaker. Authenticity seemed to be questioned in both imitators and genuine non-native speakers if typical German phonetic/phonological patterns are present, i.e. have been successfully acquired by the non-natives or unsuccessfully concealed by the imitators. Features identified to be crucial for accent authenticity judgements of French and American English accents seemed to be the realisation of /h/, postvocalic /r/, junctural glottalisation and realisation of final /-ən/. So, for instance, it was suggested that a full realisation of final /-ən/ seems to lead to the judgement “authentic accent” whereas accents exhibiting reduction of final /-ən/ by realising a syllabic nasal in most cases were judged to be unauthentic.

In this paper we attempt to verify the realisation of final /-ən/ as one important phonetic correlate of authenticity for a French accent in German.

The following questions are being tested:
1) Is the realisation of final /-ən/ crucial for accent authenticity judgements?
2) Does the preceding consonant affect the importance of final syllable realisation in accent authenticity judgements, i.e. is the reduction of /-ən/ in “haben” more likely to be produced by a genuine non-native speaker of German than in “Vorkehrungen”?
3) Can a manipulation of the data (e.g. elision of [ə] and inclusion of a syllabic nasal) affect the authenticity judgements?

The stimulus sentence used in the listening experiment in Neuhauser & Simpson (2007a) (Die verstärkten Sicherheitsvorkehrungen der Banken könnten Räuber abgeschreckt haben.) contains five occurrences of final /-ən/ and was analysed for twelve subjects, i.e eight native German speakers imitating a French accent and four native French speakers speaking German. These speakers belong to the following four groups of speakers which arise from the results of the listening experiment in Neuhauser & Simpson (2007a): (1) the most authentic sounding native German speakers imitating a French accent (N=4); (2) the least authentic sounding native German speakers imitating a French accent (N=4); (3) the most authentic sounding native French speakers speaking German (N=2); (4) the least authentic sounding native French speakers speaking German (N=2).

Figure 1 shows that only three out of six speakers judged to be producing authentic foreign accents reduced final /-ən/, and all of them in only one of five possible cases (in the word “haben” each). For the group of speakers judged to be producing unauthentic accents there were eleven cases of reduction of final /-ən/ which were distributed across five speakers (out of six). Interestingly the two speakers judged to be least authentic in fact were native French speakers both producing reductions of final /-ən/ not only in “haben” but also in other cases as well. These findings strongly suggest that not only may the realisation of final /-ən/ have an influence on authenticity judgements but also support the claim that the preceding consonant might be relevant as well.

An experiment is currently underway to verify the findings of this preliminary analysis. Using praat, original utterances from both authentic and non-authentic sounding speakers are manipulated in both directions, i.e. by replacing [-ən] sequences with syllabic nasals and vice versa. Different groups of listeners are being asked to judge the accent authenticity of the original and the manipulated stimuli.
**Figure 1** Number of reductions of final syllable (i.e. /-ən/ as syllabic nasal) as a function of judged authenticity. The dotted line represents the 50% border of accent authenticity judgements.

**References**
Vowel reduction patterns in spontaneous speech

Gea de Jong¹, Toby Hudson¹, Kirsty McDougall¹ and Martin Duckworth²

¹Department of Linguistics, University of Cambridge, United kingdom
gd288|toh22|kem37@cam.ac.uk
²The College of St Mark & St John, Plymouth, UK.
mduckworth@marjon.ac.uk

When vowel formants are measured in forensic recordings, one can expect to measure values which are indicative of a slightly centralised pronunciation that is inherent to spontaneous speech. Compared to read or highly intelligible speech spontaneous speech is faster, and often contains more modified and reduced forms of consonant and vowel segments (Picheny, Durlach and Braida 1986). These processes are common across languages (Crosswhite 1999) and mostly unproblematic from a perceptual perspective: a listener has such powerful strategies at his disposal at phonological, grammatical, semantic and pragmatic levels to restore missing acoustic information, that a talker is not obliged to articulate every phoneme at its target position (Bergem and Koopmans-van Beinum 1989). Lindblom (1963) found that vowel formants were affected by the consonant context and the duration of the vowel. The magnitude of their displacement towards neighbouring consonants was related to the duration of the vowel: shorter vowels showed a higher degree of assimilation or target undershoot. Like others he also found that vowel reduction seems to affect the height contrast, or F1, before it affects the back-front contrast (Flemming 2005, Van Son and Pols 1992). In his 1990 paper, Lindblom suggests that speakers vary their articulatory output along a continuum from hypo- to hyperspeech.

The present study looks in detail at the different vowel reduction patterns which individuals exhibit in different speaking styles and contexts and their implications for forensic phonetic analysis. On average what degree of reduction can one expect? Do most speakers reduce their vowels to a similar degree? Can speakers be categorised in one of two groups (severe vowel reducers vs mild vowel reducers) or does a more gradual pattern exist?

Read speech recordings of male speakers of Standard Southern British English (SSBE) aged 18-25 from the Dynamic Variability in Speech (DyViS) corpus are analysed (University of Cambridge: UK ESRC RES-000-23-1248). F1 and F2 frequencies of vowels in read speech are compared with those in spontaneous speech in two different forensically relevant contexts: 1) a simulated police interview and 2) a more relaxed telephone conversation with the ‘accomplice’. Finally, now that the DyViS database is complete, F1 and F2 frequencies of /i, æ, ɑ, ɔ, ʊ, u/ in hVd contexts are presented for all 100 SSBE speakers of the DyViS corpus.

References


Formant frequency measurements become more and more important in forensic casework and have been shown to contain important speaker-specific information. A simple and relatively new method of formant analysis is the long-term formant distribution (LTF). This method involves pooling formant values of all vowels of a speaker, leading to only one average value and one standard deviation per formant. For each speaker there is an individual distribution curve as can be seen in Figure 1. Nolan & Grigoras (2005) introduced this method and provided an insightful comparison of LTF’s with long-term average spectra and formant measurements of individual vowels – working out the particular advantages of the LTF method. In order to be able to use LTF in a Bayesian approach (i.e. for the Likelihood-Ratio addressing the similarity and the typicality aspect), it is necessary to have an LTF reference database of the population. Whether or not Bayesian statistics are applied, having a good idea of the distribution of long term formant frequencies in the population is important in all types of forensic casework where the LTF method is used.

This work presents a database of LTF values for spontaneous and read speech of 71 German male adult speakers recorded via mobile phone microphones. The data was taken from the speech corpus “Pool2010” provided by the Bundeskriminalamt (Jessen et al., 2005). In order to obtain the LTF values, each sound file was treated as follows: All non-vowel sounds (and silences) were cut out of the signal so that only a sequence of vowels remained as speech signal. An LPC-based automatic formant tracking was applied by the programme WaveSurfer, which was manually corrected. For each resulting formant measurement file, LTF and other analyses were performed.

Most speaker specific seems to be the LTF value of F3 (=LTF3) as there are the largest between-speaker differences. Further, there are significant differences between the values of spontaneous and read speech; LTF values of read speech are higher. Accordingly one should be careful comparing LTF’s of read and spontaneous speech. Figure 2 shows the differences of LTF2 and LTF3 between our subjects (along the abscissa) in ascending order of LTF3 values and the differences of spontaneous and read speech within each subject (black vs. grey). Expectedly there are only subtle differences in LTF1. As can be seen there is no correlation between LTF2 and LTF3; LTF3 is ascending continuously while LTF2 goes zigzag. The slightly higher position of read speech (grey) compared to spontaneous speech (black) supports the finding of read speech being higher.

**Figure 1** Frequency of occurrence of formant values measured 100 times a second in speech samples of two speakers. Black = F1, light grey = F2, dark grey = F3.
Another question to be investigated was whether there is a net duration threshold value of available speech material beyond which LTF’s are saturated. It was found that such a threshold can be placed at around five to eight seconds of pure vocalic stream, depending on the formant and the speaking condition. Looking at single speakers, however, it was found that the threshold can vary between speakers considerably.

The presentation of this work aims at showing both general findings about LTF and single-speaker patterns and comparisons. Hence, averaged results as well as speaker comparisons will be presented.

References
A Method for Reducing Formant Measurement Errors in Synthetic Speech

Philip Harrison
J P French Associates, York, UK &
Department of Language & Linguistic Science, University of York, York, UK
pth@jpfrench.com

Previous work by the author has concentrated on formant measurement variation within and between different software packages used by forensic phoneticians (Harrison, 2004, 2005 & 2006) and more recently formant measurement errors obtained from synthetic speech (Harrison, 2007). This presentation introduces the latest development in this ongoing work, namely a method to reduce the error in formant measurements for synthetic speech.

The current work is based on the analysis of speech produced by the source-filter synthesis method where formant frequencies are specified during the synthesis process. By plotting the specified values against the resulting measured formant values obtained from an LPC formant tracker it is possible to visualise the relationship between these two sets of formant values. If the equation that represents this relationship is derived (by using Fourier series analysis) then it is possible to determine from a measured formant value what the original specified formant value was. In other words one can recover the specified formant value, effectively removing any measurement error.

However, there are two complicating factors. Firstly, the relationship between the specified values and measured values is dependent on the fundamental frequency. So in order to obtain the specified value the f0 must also be measured. This process is not perfectly accurate so any errors in the pitch measurement will affect the accuracy of the effective error removal for the formant values. Secondly, there are regions within the vowel space where the specified value cannot be recovered since a many-to-one mapping exists where a single measured formant value could have originated from one of either two or three specified values. The size and occurrence of such regions within the vowel space is dependent on both the LPC order of the formant tracker and the fundamental frequency.

The process described above will be demonstrated and the resulting corrected formant values presented. The applicability to real speech data will also be considered.

References


1. Introduction

Understanding the influence of stress on speech is a major concern in forensic speaker identification. A problem in the past has been the lack of comparibility of studies. This is due to diverging definitions of stress and that the effect of stress on speech has been the object of interdisciplinary research (Benson 1995, Jessen 2006, Streeter 1983). Therefore, the stress specification model developed on the basis of the results of the ESCA-NATO Workshop “Speech under Stress” was applied (Murray 1996). This study examines the effect of two stressors, situational and cognitive (Hicks 1979), on selected vocal parameters.

2. Method

The stressors were generated according to the “Trier Social Stress Test – TSST” (Kirschbaum 1993) which has shown to produce significant physiological responses under laboratory conditions. Speech recordings were made of 15 male and 15 female subjects under two stress conditions, simulated job interview (situational) and arithmetic problem solving (cognitive). These conditions are to be classified as 3rd-order stressors. Fundamental frequency data (mean, median, standard deviation) as well as formant history (F1, F2, F3) were extracted using Praat (Boersma 2006). Syllable and articulation rate were determined manually.

3. Results

Frequencies of F1 and F3 under cognitive stress were found to be higher than under situational stress. Syllable and articulation rate, on the other hand, were both higher under situational stress. No significant differences were determined relative to fundamental frequency parameters. Gender-related responses were also not found.

References


A voice parade was conducted in a case involving a purported earwitness identification. The procedure was carried out at the instruction of the trial judge in a case in progress, where the judgement of an earwitness, purporting to identify the perpetrator of a serious assault, was called into question by defence counsel.

**Procedure and Outcome**

The procedure followed was that laid down in the *UK Home Office Circular 057/2003*, and described in Nolan (2003). Trials were conducted with panels of mock witnesses unfamiliar with the voice of the suspect. Preparation of the materials was hampered by the very poor quality of the recording of the suspect, and measures were adopted to equalize the noise conditions across the suspect’s sample and the accompanying foils by the addition to the foil samples of noise matching that observed in the suspect’s recording. Noise was added variously to the different foil samples at random levels of amplitude above and below the level observed in the suspect sample, to avoid creating uniformity in the foil samples, against which greater variability in the background noise in the foil sample might have proved distinctive.

In a first round of trials a significant proportion of the mock witnesses successfully identified the suspect’s voice. Mock witnesses who successfully identified the suspect sample were asked to account for their choice: while one suggested that the background noise in the suspect sample was qualitatively different from the foil samples, most witnesses identified the subject matter of the suspect’s recording as critical to their judgement. The case involved a violent attack by masked intruders to a woman’s house; in keeping with the guidelines, mock witnesses were made aware of this background to the case. The suspect sample featured repeated references to *she* and *her*; while these were unrelated to the case in question, listeners took these references to relate to the victim of the assault.

In the light of the unsuccessful outcome to the first round of trials, a further reference sample of the suspect was obtained, by order of the trial judge and after protracted legal argument. This further sample was significantly better in quality than the original recording, and necessitated the reverse process from that carried out in the first round: background noise in the original foil samples (which were all taken from authentic police interviews with suspects of unrelated crimes) needed to be reduced in order to match the quality of the new recording. The subject matter of the new recording concerned the suspect’s domestic routine, with no content which might have been erroneously related to the crime under investigation. A second panel of listeners, different from the first, were unable to identify the suspect from among the foils. It may be speculated that the domestic nature of the subject matter mitigated the likelihood of mock witnesses linking the suspect to the crime in question.

Materials were prepared and the actual parade procedure conducted, in keeping with the prescribed guidelines. The outcome was that the witness failed to identify the suspect, despite the fact that the witness had claimed to recognize the voice of her attacker at the time of the assault.

**Implications**

The case bears significant resemblances to the case described in Nolan (2003) in particular in the use of a voice parade in a case where the witness claimed to have recognized the voice of the suspect at the time of the incident. A nascent trend may be discerned, in that courts and police forces may view the voice parade procedure not as an exercise in identification parallel to that of an eye-
witness lineup, but as a measure by which to verify a witness’s claimed ability to identify a known speaker by voice alone. The suitability and efficacy of the voice parade as a measure of this kind is open to question.

References
Voice Recognition within a closed set of family members

Elizabeth McClelland
Forensic Analysis of Language Samples, Edinburgh, UK
earsemc@googlemail.com

The study was designed as a follow-up to research carried out in 2000 which tested the accuracy of voice recognition within a family set of 14 speakers aged 17 - 89, including one set of male twins. The results indicated that certain voices were more difficult than others to recognize and that the ability to recognize known voices accurately over the phone varied from individual to individual. The voices in the test line-up that were most prone to misrecognition were those of the younger age group. The eight cousins, now aged 25-33, who were part of the earlier study were re-recorded speaking a rehearsed, linguistically neutral message onto a telephone answering machine. The panel of subjects included the speakers whose voices been most difficult to recognize in the original study.

The partners of each of the eight subjects were also exposed to the same compilation of stimuli and asked to indicate which voice was that of their partner. Three out of the eight failed consistently to recognize their partner’s voice.

Test Methodology

A compilation recording was prepared containing three line-ups, each consisting of ten stimulus utterances extracted from the recorded messages. The three line-ups included the voices all eight subjects, with two voices repeated.

The line-ups were played once to the subjects. They were told that the same voice might occur more than once in each line-up and that each set contained only their own voice and the voices of their cousins.

The accents of the subjects in the line-ups can all be informally categorized as ‘posh Scottish’. The relationship between difficulty of voice recognition and phonetic markers in the subjects’ voices was far from obvious.

Analysis and Results

The same three male and two female voices that had proved most challenging in the earlier test continued to be mis-identified.

Subjects who achieved high scores in identifying voices accurately in the first study again showed proficiency in the current test.

Analysis of the speech data produced the following results:

(i) Voice quality profiles, speaking rate, rhythm and timing in the mis-identified male speakers were similar. The two challenging female voices also shared similarities in terms of the same set of features.

(ii) Certain vowel realizations differed both between the three male speakers and the two female speakers whose voices were most frequently confused. The differences did not appear to assist the subjects in distinguishing the voices.
(iii) The three male voices that were most frequently confused with each other had very similar pitch measurements. However, the voice of the twin brother of one of subjects in this group – whose pitch was also very similar - was consistently correctly identified. Conversely, pitch measurements in the two female voices which were confused exhibited a clear difference.

(iii)

Our test substantiated the finding of Foulkes (2000) that subjects whose F0 were at the extremes of the measurements for their gender in the test group were consistently correctly identified.

References


All Ears – An earwitness research programme

Anders Eriksson¹, and Lisa Öhman²

¹Department of Linguistics, University of Gothenburg, Sweden
anders.eriksson@ling.gu.se

²Department of Psychology, University of Gothenburg, Sweden
lisa.ohman@psy.gu.se

The goal of the research programme is to investigate the reliability of earwitnesses under a variety of conditions. It is a three-year programme which began in the autumn of 2007 and is funded by the Crime Victim Compensation and Support Authority (Grant 03347/2007). An equal number of male and female speakers will be used in all studies. With the exception of Study 3, speakers and speech material will be the same. An important consideration has been to make the recordings and perform the tests in such a way that no advanced technical resources are required to set up a comparable test. Recordings were therefore not made in a professional sound studio, but in a reasonably quiet and echo free room. The recorder was a solid state recorder with built-in microphones. Sound quality is good, but not studio quality. In order to obtain identical speech material for the telephone recordings used in Study 2, the speech material was simultaneously recorded over the telephone. The selection of the “perpetrator” was decided by the authors. The selection of the foils for the line-ups was done with the help of a perceptions test. The latency time will be 2 weeks in all studies.

Study 1. Reliability as a function of listener age.

Speaker recognition for two groups of children, aged 7–8 years and 12–13 years, will be compared to that of a group of adult listeners. Each group consists of 80 participants, 40 of which will be presented with target present line-ups and the rest with target absent line-ups. The children are school children and the tests are carried out in their respective schools. The adult listeners are tested at the departments of linguistics or psychology. At the time of writing this abstract, most of the tests have been completed. Preliminary results will be presented at the conference.

Study 2. Reliability as a function of presentation format.

There is an ongoing debate about whether telephone recordings should be used for the line-ups in cases where the witness has heard the perpetrator over the telephone, but little is actually known about it and what is known indicates that it may not make a difference. In Study 2, this question will be addressed. 160 adult subjects will be used as listeners, 80 of them will be presented with telephone recordings of the perpetrator’s voice and 80 with direct recordings. In the line-ups half of the participants in each group will be presented with direct recorded line-ups and the other half with telephone recorded ones. Only target present line-ups will be used.

Study 3. Reliability as a function of the number of heard speakers.

It is not uncommon, for example in kidnap cases that several perpetrators are involved who can be heard but not seen by the victim. In a well known case in Sweden, the victim was kept in a wooden box from which he could hear, but not see, his kidnappers. In Study 3 we will investigate the effect on reliability, if any, of hearing more than one voice. 100 subjects will participate. Half of them will be presented with only one perpetrator voice and the other half will hear the chosen perpetrator together with two other voices in a conversation. In the target present line-ups the target voice will be the same for both groups.

Study 4. Reliability as a function of instructions to the witness.

Here we will investigate to what extent the reliability of a witness may be affected by instructions given in connection with the line-up. A technique called cognitive interview, which has been used in eyewitness studies will be used. 160 subjects, 80 children aged 7–8 years and 80 adults, will be used. Half of the participants will be prepared for the line-up using the cognitive interview technique while the rest will be tested as in the previous experiments.
Population statistics on the phonology of Scottish English, of which there are approximately 5 million speakers, are as yet scarce. So as to address this gap a corpus of recordings of the varieties of Scottish English spoken in Scotland’s four major cities (Glasgow, Edinburgh, Aberdeen and Dundee) was collected in 2005-7 with the assistance of an IAFPA research grant. Both sexes and a range of ages (teenagers to individuals of retirement age) are represented in the sample. In view of the intended forensic utility of the material a bias towards younger talkers was thought desirable, although the male/female split is approximately equal. Talkers were asked to read a phonetically-balanced word list (twice), two text passages, and to talk informally for a short period about a topic that interested them.

The statistics presented here are the mean and standard deviations of fundamental frequency and voice onset time (VOT) of /p t k b d g/ for each of 67 talkers from the four cities. VOT is known to be variable across regional varieties of Scottish English (Johnston 1997; Scobbie 2006; Watt & Yurkova 2007, Stuart-Smith 2008) but by and large this is based upon impressionistic observations rather than quantitative measurements. The variability among the four varieties - and between these Scottish English accents and Southern Standard British English (RP) - with respect to fundamental frequency and VOT is discussed, as is the relevance of these data to forensic casework undertaken on behalf of police forces in Scotland.

References
How speaker idiosyncratic is measurable speech rhythm?

Volker Dellwo\textsuperscript{1} and Jacques Koreman\textsuperscript{2}

\textsuperscript{1} Division of Psychology and Language Sciences, UCL, UK. 
\texttt{v.dellwo@ucl.ac.uk}

\textsuperscript{2} Department of Language and Communication Studies, NTNU, Norway. 
\texttt{jacques.koreman@hf.ntnu.no}

Previous work has demonstrated ways to measure aspects of speech rhythm on the basis of durational characteristics of consonantal and vocalic intervals (Ramus et al., 1999, Grabe & Low, 2002). While it remains a matter of ongoing debate to which degree these measures reflect perceptually salient features of speech rhythm, the present research investigates whether such acoustic measurements can contain speaker idiosyncratic information. In an experiment ten speakers of German read a short text (seven sentences) five times while varying their intended reading tempo for each repetition (in the following order: normal, slow, slower, fast, fastest possible). The rhythm measures $\%V$ and $\Delta C$ (percentage over which speech is vocalic and standard deviation of consonantal interval durations; see Ramus et al., 1999) and the vocalic and consonantal PVI (pairwise variability index, a measure of average consonantal and vocalic interval differences across speech; see Grabe & Low, 2002), as well as rate normalised versions of these measures (e.g. Dellwo, 2006), were calculated for this data. Figure 1 plots the results for $\%V$ for each speaker and intended tempo version. It can be seen that $\%V$ can vary strongly between speakers but is typically rather stable across the different intended tempo conditions within each speaker. Similar degrees of speaker idiosyncratic information could be detected in case of the other rhythm measures $\Delta C$ and PVI.

![Figure 1](image)

Figure 1 Box plot showing the distribution of the rhythm measure $\%V$ (percentage over which speech is vocalic) for 5 intended speech tempo versions of 10 speakers of German (each plot is the result of 7 measurement points, one measurement per sentence).

The results imply that the speech signal contains speaker idiosyncratic durational information which typically remain stable across drastic changes of speech tempo. Other parameters, such as fundamental frequency which have been used widely in forensic speaker identification can be demonstrated to vary significantly within speakers under such conditions. This not seldom poses problems for the identification process when speech samples of acoustic trace and comparison material in a forensic context vary in rate (for example, as a result of different emotional states the speaker/s was/were in during the different recordings).

We are currently running further experiments to investigate how measures of speech rhythm vary as a function of other prosodic parameters such as changes in fundamental frequency or inten-
We are also testing whether measurable speech rhythm remains constant for speakers across various paralinguistic conditions of speaker state (e.g. different emotions) and ultimately we are testing the effects of voice disguise on measurable rhythm. In addition to this we are looking at other ways to measure durational characteristics in the speech signal since the above discussed 'rhythm measures' may not be the ultimate parameters to capture speaker idiosyncratic information (it has been demonstrated, for example, that they also reveal language specific durational characteristics to some degree). Besides an evaluation in terms of rhythm measures, we shall also present first results of a phonetic analysis of the reduction behavior of specific speakers. In particular, differences in the intended (“canonical”) and realized speech rates (Koreman, 2006), which affect the rhythm measures, will be considered both across the utterance as a whole and for the accented and unaccented parts of the utterances separately, to evaluate possibly different speaker strategies. If speech durational characteristics remain to reveal speaker idiosyncrasies across a number of other conditions, we believe that they may be powerful parameters for speaker identification and verification purposes in the future.

References


The aim of the study is to investigate whether naïve listeners can recognise a voice spoken in a language which they cannot understand in a voice lineup spoken in a language which they can understand. 30 British naïve listeners were asked to identify a voice from seven L1 Korean speakers’ speaking English immediately after they were exposed to a voice speaking Korean. A control group of Korean listeners performed the same task. In addition, Speaking Fundamental Frequencies, Articulation Rates, Vowel Formant Frequencies, and High Formant Frequencies were acoustically analysed in order to find what parameters might have contributed to voice recognition.

In auditory-perceptual analysis with British listeners, only 27% of the listeners were able to correctly recognise the original voice and one foil was selected by 43% of the listeners. On the other hand, in auditory-perceptual analysis with Korean listeners, only 17% of the listeners had the correct answer and one foil, different from the one British subjects chose, was selected by 50% of the listeners. In acoustic analysis, it was found that there were non-significant differences in the Speaking Fundamental Frequencies of the speakers when speaking English and Korean. It is very similar for the Articulation Rates between the original voice and the one of the favoured foils. In conclusion, differences in High Formant Frequency values appeared to correlate with recognition by British listeners. Therefore, it may be no more difficult for naive listeners to recognise a voice spoken in an unknown language than in a known language. Indeed, the results from this exploratory study indicate it may even be easier. It is suggested that this method should be further developed with other speech data of various languages.

**Figure 1** The percentages of positive responses of each voice to the target voice (E6) by British and Korean listeners.
References


Population data for English spoken in England: A modest first step

Frantz Clermont\textsuperscript{1}, Peter French\textsuperscript{2}, Philip Harrison\textsuperscript{3} and Samantha Simpson\textsuperscript{4}
\textsuperscript{1,2,3}J.P. French Associates & University of York, UK
\{frantz.clermont|jpf|pth\}@jpfrench.com
\textsuperscript{4}University of York, UK
ss599@york.ac.uk

With the aim of making a beginning at establishing background population statistics for use in forensic speaker comparison cases, this study takes a subset of 25 young men (age 18-25 years) from the University of Cambridge DyVis corpus engaging in unscripted conversation over telephone lines. The accent range represented in these recordings has been described as “Standard Southern British English” (Nolan et al., 2006). A problem for projects of this type concerns the assembly of population data which have maximum generalisability. In other words, the target is to identify those features that are subject to least social and regional variation.

To this end, our initial focus has been on the six phonologically-short vowels, namely those represented by the lexical set headwords KIT, DRESS, TRAP, LOT, FOOT and STRUT (Wells, 1982). Whilst there is undoubtedly some regional and social variation (e.g., raising of the KIT vowel for Midlands English, lack of FOOT-STRUT split in Northern English accents), the degree of variation within the short vowels of English spoken in England\textsuperscript{1} is considered to be much less than that for the phonologically-long vowels and for diphthongs.

The study progressed via the following steps:

(1) All tokens of the six vowels occurring in stressed syllables were identified and marked in the sound file for each speaker;
(2) All tokens were examined spectrographically, and nuclei points were established for the taking of formant-frequency measurements;
(3) Formant frequencies (F1, F2, and F3) were estimated using an “intelligent” formant tracker (Clermont et al., 2007; Clermont, 1992), and checked against spectrograms;
(4) The following information was derived from the measured formant frequencies:
   (a) The region of vowel space (spanned by F1 and F2) utilised by each speaker in the production of each vowel within the set of six;
   (b) The degree of within- and between-speaker variation in respect of (a) above;
   (c) The degree of within- and between-speaker variation in F3, and the extent to which F3 values vary across the 6 vowel categories.

The patterns found in this corpus are discussed in terms of their relevance to the forensic speaker comparison task, and directions for an extension to the project using real case data are considered.

References


\textsuperscript{1}This claim does not apply to Scottish, Welsh, Northern Irish and certain minority accents.
Forensic Automatic Speaker Recognition – Towards Biometric Evidence of Voice

Andrzej Drygajlo
Speech Processing and Biometrics Group, EPFL, Lausanne, Switzerland
andrzej.drygajlo@epfl.ch

Forensic speaker recognition is the process of determining if a specific individual (suspected speaker) is the source of a questioned voice recording (trace). The forensic application of speaker recognition technology is one of the most controversial issues within the wide community of researchers, experts and police workers. This is mainly due to the fact that very different methods are applied in this area by phoneticians, engineers, lawyers, psychologists and investigators. The approaches commonly used for speaker recognition by forensic experts include the aural-perceptual, the auditory-instrumental, and the automatic methods. The forensic expert’s role is to testify to the worth of the evidence by using, if possible a quantitative measure of this worth. It is up to other people (the judge and/or the jury) to use this information as an aid to their deliberations and decision. This essay aims at presenting forensic automatic speaker recognition (FASR) methods that provide a coherent way of quantifying and presenting recorded voice as scientific evidence. In such methods, the evidence consists of the quantified degree of similarity between speaker-dependent features extracted from the trace and speaker-dependent features extracted from recorded speech of a suspect. The interpretation of recorded voice as evidence in the forensic context presents particular challenges, including within-speaker (within-source) variability, between-speakers (between-sources) variability, and differences in recording sessions conditions. Consequently, FASR methods must provide a probabilistic evaluation which gives the court an indication of the strength of the evidence given the estimated within-source, between-sources and between-session variabilities.

We discuss some important aspects of forensic speaker recognition, focusing on the necessary statistical-probabilistic framework for both quantifying and interpreting recorded voice as scientific evidence. Methodological guidelines for the calculation of the evidence, its strength and the evaluation of this strength under operating conditions of the casework are presented. As an example, an automatic method using the Gaussian mixture models (GMMs) and the Bayesian interpretation (BI) framework were implemented for the forensic speaker recognition task. The BI method represents neither speaker verification nor speaker identification. These two recognition techniques cannot be used for the task, since categorical, absolute and deterministic conclusions about the identity of source of evidential traces are logically untenable because of the inductive nature of the process of the inference of identity. This method, using a likelihood ratio to indicate the strength of the evidence of the questioned recording, measures how this recording of voice scores for the suspected speaker model, compared to relevant non-suspect speaker models.

It became obvious that particular effort is needed in the trans-disciplinary domain of adaptation of the state-of-the-art speech recognition techniques to real-world environmental conditions for forensic speaker recognition. The future methods to be developed should combine the advantages of automatic signal processing and pattern recognition objectivity with the methodological transparency solicited in forensic investigations.

References

The UK Position Statement on Forensic Speaker Comparison
A Response

Phil Rose & Geoffrey Stewart Morrison
School of Language Studies, Arts, Australian National University
philip.rose@anu.edu.au
geoff.morrison@anu.edu.au

A recent issue of Speech Language and the Law contained a “position statement concerning the use of impressionistic likelihood terms in forensic speaker comparison cases” (French & Harrison 2007). This position statement was the result of collaboration between a number of researchers and forensic practitioners working in the United Kingdom. It is claimed (ibid: 138) that “the new framework is, at a conceptual level, identical to that used nowadays in the presentation of DNA evidence”.

The current paper is a draft version of the authors’ response to the UK position. The authors are now collaborating with a number of researchers and forensic practitioners around the world – including automatic forensic speaker recognition – and hope to be able to include at least some of their responses. We hope that it will lead to additional discussion and refinement, and perhaps greater consensus in what is yet, unfortunately, a methodologically rather heterogeneous field (Cambier-Langeveld 2007).

We applaud the motivation of the UK position statement, and welcome its general direction. Nevertheless, of course, we have some criticism and some constructive suggestions for improvement. In particular we examine the claim, quoted above, of conceptual equivalence between the position statement and what is now recognized as the logically correct way of evaluating forensic identification evidence as exemplified in DNA profiling (Balding 2005, Gonzalez-Rodriguez et al. 2007). Other concerns involve the importance of being explicit as to how far Forensic Speaker Comparison is able to meet the scientific requirements of falsifiability and replicability as articulated in Daubert (1993); of distinguishing between categoricity and gradience in traditional FSR features; of showing how to relate the proposed assessments of distinctiveness and consistency - the latter of which appears to be particularly epistemologically weak (Robertson & Vignaux 1995); and of acquiring appropriate reference data (Rose 2007).

References
Daubert (1993) Daubert vs Merrell Dow Pharmaceuticals, Inc. 113 S Ct 2786.
Will we be really able to estimate evidence in this scenario?

Matteo Nunziati
Dipartimento di Meccanica e Tecnologie Industriali (DMTI), University of Florence, Florence, Italy
matteo.nunziati@unifi.it

Many computational and analytical models are employed by forensic scientists to assess the problem of speaker recognition. Model performance depends on a number of factors which can be divided in both linguistic and technical ones. Linguistic factors account for issues related to the spoken language/dialect and kind of speech acquired during a recording session (e.g. the well known Lombard effect). Technical factors take into account the signal quality and the amount of available features which have been extracted from a recording.

An additional problem for many speaker recognition researchers is induced by model porting practices. It is common to perform a recognition experiment employing models which have been originally developed for different tasks. A remarkable example is the wide employment of the multivariate kernel density model (MKV) published by Aitken and Lucy (2004). Such a model has been developed to estimate evidence involving glass fragments or paint stains. It lacks of any between-session-variability modelling, nonetheless it is commonly adopted by many scientists.

Even if solutions to many of these issues are still undisclosed, novel procedures have been developed to assess the relevance of such lacks. Noticeably, the FoCal toolkit described in Brummer and du Preez (2006) seems to be a really effective tool for the evaluation of discrimination and calibration losses induced by limited knowledge and modelling. The FoCal toolkit is able to compensate many model lacks and to fuse two or more responses together. This feature is based on a data-driven approach, thus no specific knowledge of the problem has to be provided. As a consequence, we can improve the overall performance of any simplified model by simulating different scenarios and tuning responses, according to expected real relations between test recordings.

A recent paper by Gonzalez-Rodriguez et alii (2007) has shown that the FoCal toolkit is an effective instrument to provide a scientific and transparent (according to well known Daubert rules) approach to the problem of system goodness assessment. This approach has been followed by other authors, including the new manuscript by Morrison and Kinoshita (2008) (currently under review; a preview of the paper is kindly distributed on line by authors).

Even though the FoCal toolkit is a remarkable instrument, it is limited in that it introduces an information-theoretical approach only for specific prior odds conditions. As information theory seems to be the correct approach to evaluate a recognition system, a solution has been recently provided via the so called ECE plots by Ramos et alii (2007).

In this work, we describe how both the FoCal toolkit and the ECE plot can be jointly employed in a real prosecution to predict recognition accuracy. Such pre-analysis can be adopted as a scientific and transparent approach to accept or reject a recording.

At first, we propose to demonstrate the accuracy of a certain technology in standard conditions (such as clean/low noise GSM recordings), thus no doubt can be arisen in relation to system capabilities. Later, we show how specifically developed tests can demonstrate the degradation of accuracy caused by limited quality recordings. Specifically, it is possible to demonstrate the accuracy of a recognition conducted on the basis of one rather than two or more vowels. Additionally a demonstration of misdetection is provided, simulating the same kind of noise detected in some real recordings.

As ECE plots interpretation involves many non trivial concepts related to statistic and information theory, we use this paper to also provide a simplified version of the plot. The proposed solution, called here k-plot, is meant as a preliminary experiment to balance both court requirements of transparency and easy interpretation.
References


The influence of intra-speaker variability in automatic speaker verification using F0 features

Timo Becker
Acoustics Research Institute of the Austrian Academy of Sciences, Austria
timo.becker@oeaw.ac.at

Recently, F0 features have been used more frequently in research concerning speaker recognition (for example Kinoshita et. al. 2007). While it is important to investigate all available speaker characteristics, the intra-speaker variability has to be known to avoid misclassifications. Also, the robustness of F0 features and classification methods that depend on those features have to be determined. In forensic casework, mismatch in speaking styles occurs frequently between offender and suspect recordings. Here, the influence of mismatch in speaking styles concerning F0 is investigated.

Automatic speaker verification using F0 features comes along with the problem of feature correlation. Here, a new method using Principal Components Analysis (PCA) is used for the decorrelation of features. A separate corpus is used to estimate the new feature space basis on which the training and test corpus data is projected. Additionally, the dimensionality can be reduced without information loss (Becker & Kreuzer 2007).

290 male speakers from the TIMIT corpus were used to estimate the new basis by application of PCA. The ‘Pool 2010’ corpus recorded by the Department of Speaker Identification and Audio Analysis, Bundeskriminalamt, Germany (Jessen et al. 2005) was used for the experiment. Recordings of reading (R) vs. spontaneous (S) and free (F) vs. Lombard (L) speech were used for speaker verification tasks. The recordings were halved to create test and training sets. The resulting average voiced signal length was 10 seconds for read speech and 29 seconds for spontaneous speech.

F0 values were extracted automatically with STx (www.kfs.oeaw.ac.at), based on the method described by Boersma (Boersma 1993). Feature vectors consisted of the six features arithmetic mean, standard deviation, variation coefficient (Jessen et al. 2005), base line (Fb) (Lindh & Eriksson 2007), Skewness and Kurtosis. Dissimilarity was expressed in Euclidean distance in the feature space. Both raw feature vectors and two-dimensional transformed vectors after application of PCA were used.

Equal error rates (EERs) without speaking style mismatches ranged from 15% (RF) to 23% (SL) for raw vectors, and from 10% (RF) to 20% (SL) for PCA transformed vectors. GSM transmission showed only slight differences of EERs. Performances of speaking style mismatches for the GSM transmitted recordings showed the best results for an R/S mismatch of free speech (30% raw, 26% PCA) and an R/S mismatch of lombard speech (32% raw, 29% PCA). All F/L mismatches (including double mismatches where both R/S and F/L mismatched) showed very low performances (47% to 49% raw, 38% to 42% PCA).

These results are supported by the results from Braun (1995) and Jessen et al. (2005), who found differences of F0 mean and F0 deviations in different speaking styles, as well as speaker individual variations (Jessen et al. 2005).

Summarising, Lombard speech leads to low EERs, especially in combination with speaking style mismatch. An R/S mismatch leads to performance loss as well, but to a lesser degree. The PCA method generally improves performance. However, due to the performance degradation caused by speaking style mismatches, the usage of F0 as a speaker specific feature in forensic casework is questionable whenever a mismatch occurs.

References


Robustness of Forced Alignment in a Forensic Context

Jonas Lindh
Department of Linguistics, University of Gothenburg, Sweden
jonas.lindh@ling.gu.se

Sometimes in forensic phonetic casework a text file with an orthographic transcription is accompanying the audio file/s. In such cases, automatic forced alignment can be used. Forced alignment is a technique based on speech recognition technology, where phones are recognized, labeled and time aligned with an audio signal. The technique has been used for a long time in speech technology and evaluated is considered reasonably robust for high quality recordings. In the present study, the robustness of forced alignment as a function of audio quality is tested, with or without the help of orthographic transcription. The results indicate that it is possible to use the technique at least as a complement and semi-automatic help in forensic phonetic casework.

Forced Alignment

When concatenation speech synthesis became popular as a technique and large databases for labeled speech were used as a basis for concatenation, researchers became aware of how time consuming manual labeling of all the recordings can be. To be able to speed up the process methods for automatic labeling and segmenting were sought for. The most successful attempts that followed were based on automatic speech recognition techniques such as Dynamic Time Warping (Malfrère et al., 1998) and Hidden Markov Models (Brugnara et al., 1993). The early results show successful alignment of up to 80% compared to manually labeled data (Hosom, 2000). It is not obvious how to quantify alignment accuracy, but a usual method is to consider the manually labeled data as the gold standard and consider automatic alignment to be correct when the deviation from manual alignment is within 20 ms.

Method

Nalign is an aligner using Mel Frequency Cepstral Coefficients (with delta and acceleration coefficients), Hidden Markov Models and Viterbi recognition to generate phone and word labels (Sjölander, 2001). On high quality speech it performed 85.1% correct labeling in one study (Sjölander, 2003) and between 90-95% in another study (Sjölander and Heldner, 2004) compared to manual labeling. The aligner recognition was trained on silence and so called garbage models to handle spontaneous pauses and other disfluencies.

Together with praat scripts (Boersma and Weenink, 2008) the tool was tested on a small corpus of simultaneously recorded material of different qualities (Livijn, 2004). “The north wind and the sun” was recorded in Swedish using one AKG SE-300B condenser microphone (reference recording - REF), one AKG D3700S dynamic microphone (to cassette recorder - CAS), a Panasonic micro cassette recorder (MCA) and a Sony Ericsson T610 mobile phone (MOB) simultaneously. The recording took place in the anechoic chamber at Stockholm University. The condenser microphone was connected directly to the sound card of a PC and recordings were made using a sampling rate of 44.1 kHz and 16 bit amplitude resolution. The second microphone was connected to a Sony TC-D5M Cassette deck (cassette - CAS). For this pilot experiment only the very first sentence was used together with a preceding “OK”.

All four recordings were aligned twice. The first automatic alignment was made using phone recognition, i.e. without any orthographic text or linguistic input but solely based on the trained vectors. The second alignment was made with an orthographic transcription.

Results

The results show a decrease in correct alignment with the decrease in audio quality as expected. In the present study only the phone level alignment with an orthographic string has been considered.
Table 1. The different audio qualities compared to the manually labeled (Gold standard).

<table>
<thead>
<tr>
<th>Audio</th>
<th>% correct vs Gold standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>REF</td>
<td>91.30</td>
</tr>
<tr>
<td>CAS</td>
<td>94.20</td>
</tr>
<tr>
<td>MOB</td>
<td>78.26</td>
</tr>
<tr>
<td>MCA</td>
<td>72.46</td>
</tr>
</tbody>
</table>

It should be pointed out that both MCA and MOB at different places introduced an extra silence (one for [s] in MOB and in MCA this occurred at the place of a glottal stop). No other incorrect labels were noticed in the comparison.

It is also somewhat worth noting that the CAS recording produced better alignment than the reference recording. The results are similar to the results presented in earlier work (Sjölander 2001; 2003).

**Future Work**

The experiment gives an idea about how well an automatic aligner can perform for material with different audio qualities. But more experiments are needed to get a more precise picture of how the correctness of automatic alignment varies as a function of audio quality in comparison to manual labeling. The method should also be tested on authentic forensic data.

**References**


Interest and limit of an open source toolkit for speaker recognition

Jean-François Bonastre
University of Avignon, LIA, Avignon, France
jean-francois.bonastre@univ-avignon.fr

In this paper, the author aims to present two main points, the ALIZE/SpkDet open source software and the evolution of the performance of such a system upon the years. Finally, a third section tries to analyse this performance evolution in term of progress in the speaker recognition field.

ALIZE/SpkDet
ALIZE/SpkDet is an open source software packages for text independent speaker recognition. This software is based on the well-known Gaussian Mixture Model associated with the Universal Background Model (UBM/GMM) approach. It includes also the latest speaker recognition developments such as Latent Factor Analysis (LFA) and unsupervised adaptation. Discriminant classifiers such as Support Vector Machines (SVM) are also provided, linked with the Nuisance Attribute Projection (NAP).

Evaluation of the performance during the last NIST SRE evaluations
Table 1 presents the evolution of the performance within the framework of NIST-SRE evaluation campaigns in terms of minimum of the detection cost function (minDCF) and equal error rate (EER) upon the years. Only the performance of the UBM-GMM system is presented but the results are similar for the SVM based systems. It shows mainly the impact of the session mismatch reduction. When a session mismatch reduction technique is used (Bayesian Factor Analysis), a drastic improvement of the performance is shown.

Table 1. Results of UBM-GMM LIA systems from 2004 until 2007 (on NIST 2005, core task, male only corpora).

<table>
<thead>
<tr>
<th>Year</th>
<th>minDCF (*100)</th>
<th>EER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004 GMM-UBM</td>
<td></td>
<td>9.64</td>
</tr>
<tr>
<td>2006 GMM-UBM</td>
<td>3.37</td>
<td>8.67</td>
</tr>
<tr>
<td>2007 GMM-SFA (GMM-UBM + Factor Analysis session mismatch reduction)</td>
<td>1.94</td>
<td>4.38</td>
</tr>
</tbody>
</table>

Table 2. Results of various LIA systems based on ALIZE/SpkDET using the unsupervised speaker model training mode, and the “oracle” (supervised) training mode (on NIST 2005, core task, male only corpora).

<table>
<thead>
<tr>
<th>Year</th>
<th>minDCF (*100)</th>
<th>EER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006 GMM-UBM (reference)</td>
<td>3.37</td>
<td>8.67</td>
</tr>
<tr>
<td>2007 GMM-SFA – standard training</td>
<td>1.94</td>
<td>4.38</td>
</tr>
<tr>
<td>2007 GMM-SFA – unsupervised training</td>
<td>0.81</td>
<td>2.27</td>
</tr>
<tr>
<td>2007 GMM-SFA – oracle training</td>
<td>0.56</td>
<td>1.71</td>
</tr>
</tbody>
</table>

Table 3. Effect of artifact-free impostor voice transformation.

<table>
<thead>
<tr>
<th></th>
<th>Baseline system</th>
<th>With transformation</th>
</tr>
</thead>
<tbody>
<tr>
<td>EER %</td>
<td>8.54</td>
<td>35.41</td>
</tr>
<tr>
<td>minDCF (*100)</td>
<td>3.58</td>
<td>9.41</td>
</tr>
<tr>
<td>False acceptance %</td>
<td>0.88</td>
<td>49.72</td>
</tr>
<tr>
<td>False reject %</td>
<td>27.45</td>
<td>27.45</td>
</tr>
</tbody>
</table>
Table 2 shows the influence of the amount of available training data in order to model a given speaker, using NIST SRE unsupervised training condition. In this condition, the test data could be used to improve the corresponding speaker model, in an unsupervised manner (the system doesn’t know if the test belongs to the speaker or not). Another time, it is clear that increasing the amount of training data is a clear factor of performance improvement.

Analysis of the performance improvement

As we have shown in this article, current speaker recognition systems are able to deal with large and increasing amounts of easy to obtain training data, either to reduce the session mismatch problem or to increase the quality of the targeted speaker models. In some previous works, we proposed to artificially transform the voice of the impostors to cheat a speaker recognition system (i.e. after the transformation, the system should recognize an impostor voice as coming from a targeted speaker). The experimental validation is synthesized in Table 3. This transparent transformation technique introduces a significant factor into the speaker recognition system, as the false acceptance rate increases from less than 1% without the transformation to about 50% when the transformation is used. Indeed, if it is possible to transform the voice of an impostor, with inaudible artificial modifications, and disrupt a speaker recognition system to such a large extent, is the information used by the system that user specific?

This fact doesn’t challenge the interest of the work done in the automatic speaker recognition area. It is clear that important progress has taken place in the last decade; for example, in the session mismatch area, which was and remains a key challenge for speaker recognition.

Conclusion

The aims of this paper are to put the focus on the danger constituted by a direct use of evaluation campaign results - based on averaged error rates - as the only criterion for evaluating the current performance and the potential of speaker recognition research and technology. It is dangerous both in terms of potential application and in terms of research orientation: there are other targets than the “performance” as measured currently. This problem is more critical in the forensic field than in the commercial area. Looking at the different points highlighted in this article, we affirm that forensic applications of speaker recognition should still be taken under a necessary need for caution. In the same time, fundamental research works on the speaker specific information should be proposed, using the new potential of automatic approaches in order to work on large databases.

Disseminating this message remains one of the most important responsibilities of speaker recognition researchers.

References


W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff (2006). SVM based speaker verification using a GMM supervector kernel and (NAP) variability compensation, in ICASSP.


Automatic Acoustical Gunshot Recognition

Anil Alexander\textsuperscript{1}, Oscar Forth\textsuperscript{1}, Cameron S. Miner\textsuperscript{2} and James Reames\textsuperscript{3}

\textsuperscript{1}Research & Development, Griffcomm Ltd, United Kingdom
\{anil|oscar\}@griffcomm.co.uk
\textsuperscript{2}Integral Design and Development, \textsuperscript{3}JBR Technologies, USA
csm\@i-d-2.com

This study addresses the issue of automatic recognition of gunshot ‘signatures’ from audio recordings made while firing weapons. Acoustic recognition of gunshots and locating their sources is useful in forensic post-event analysis, to build an accurate picture of the sequence of possibly disputed events. Live analysis of recordings could provide a tactical advantage in the protection of people or assets, etc. Famously, the analysis of the acoustics of gunshots has been applied in the John F. Kennedy assassination, concerning the provenance of shots from a second gunman. In this study, we consider applying and adapting state-of-the-art methods from automatic speaker recognition to identify the type of weapon fired based on the acoustics of the gunshot.

In automatic speaker recognition, the statistical models of acoustic parameters of the speaker’s voice and the acoustic parameters of questioned recordings are compared. The comparison of the two recordings can either be done using simple template-matching methods, or using statistical or probabilistic matching. Stochastic modelling is based on using probabilistic pattern matching, and the assumption that is tested is that the test observation and the training data were generated from the same underlying process (in this case, the same weapon). The relatively short lengths of gunshot sound impulses require the training process of these weapon signatures to be suitably adjusted. We have experimented with and adapted various state-of-the-art methods to recognize the gunshots from an audio stream. We have used a dataset of over 40 gunshots from three different weapons, along with baseline systems using template-matching algorithms as well as statistical probabilistic pattern matching.

Recordings were made of automatic, semi-automatic and bolt-action gunfire. The weapons whose signatures were compared were the Ruger (Model 77), the Glock (Model 21), and the Bushmaster M4. The recordings of shots from the Ruger (bolt action) and the Glock (semi-automatic) were made in an open field. The 9mm Bushmaster M4 (automatic and semi-automatic modes) was shot in a standard indoor shooting range, approximately 12 feet from the closest wall. The stereo recordings of these gunshots were made on high-quality recorders placed at chest level of the shooter. Both standard and high sound pressure level (SPL) microphones (Tibbetts Industries 151-01) were used, and recordings were made at 11025 Hz sampling in uncompressed, linear PCM, wave format. The files were manually segmented into individual gunshots of approximately 1 to 3 seconds in duration.

Acoustic spectral feature extraction was performed by directly applying automatic speaker recognition front-ends to the recordings. We used Mel frequency cepstral coefficients (MFCC). The MFCC-based feature extraction technique basically includes windowing the signal, applying the fast Fourier transform (FFT), taking the log of the magnitude and then warping the frequencies on a Mel scale, followed by applying the inverse FFT. The probability densities of the acoustic features were modelled using the Gaussian mixture model (GMM) [Reynolds, 1992]. The conditional probability of the test observation, given the trained model, is used as a measure of the similarity of the test and training frames.

In the results presented below, we performed a pair-wise comparison on a subset of the gunshot recordings (using five for each of the weapon type and mode combinations). The similarity scores obtained are illustrated in a score map (Fig.1). Here we observe the emergence of four distinct sets of high-similarity scores obtained, when recordings from the same type of weapon are compared. These scores are significantly higher than the similarity scores obtained from comparing recordings from different weapon types. We also observed that compared to speaker recognition, a smaller
number of Gaussian components were sufficient to adequately model the input data. The Detection Error Tradeoff (DET) plot, which is a representation that compares false acceptance and false rejection rates for a recognition system, is shown in Fig. 2. The equal error rate (EER), which refers to the threshold where false acceptance and rejection rates are equal, was 8%, using only 12 Gaussian components.

![1. Pair wise comparison score map](image1)

![2. DET Plot](image2)

**Figure 1 & 2:** 1- Results of the pair-wise comparisons of recordings from each weapon type (red denotes a higher score – blue and green denote a lower score) 2 –Detection Error Tradeoff (DET) curves for this test with an EER of 8%.

In this study the recording conditions have been relatively consistent, and the same recorder was used for all the different recordings. In practice, the recorded sound can also be affected by distances from the shooter, speed of the projectile, environmental conditions such as temperature and humidity, as well as the angle between the recorder microphone and the direction of fire [Maher, 2007]. In addition, the accuracy of recognition will depend greatly on the sensitivity, frequency response of the microphone, and the sampling rate that the recorder can support. The recording conditions, which include the recording and transmission devices used, can seriously influence the acoustic characteristics of the recorded gunshot sounds [Koenig et al, 1998]. Mismatched recording conditions between training and testing data can, as with automatic speaker recognition, adversely affect the recognition results. Mismatched conditions are often encountered in real-life forensic situations. A detailed analysis of mismatch is out of the scope of this preliminary study.

Future work would include automatically isolating short impulsive noises like gunshots, and automatically extracting and comparing distinctive spectral features of the input stream against a robustly trained database of acoustic signatures. We will consider the adaptation of models for variations in background noise, environmental conditions, as well as the adjustments for distances and angles between the recorders and the sound sources.

We would like to thank Kevin O’Neil of the Stratham Police Department, NH, USA, for his assistance in providing us with these recordings.

**References**


Effects of the telephone on perceived voice similarity

Toby Hudson, Kirsty McDougall, and Francis Nolan
Department of Linguistics, University of Cambridge, UK
toh22|kem37|fjn1@cam.ac.uk

This study concerns forensic cases where an earwitness claims to be able to remember and recognise a criminal’s voice heard over the telephone. The accuracy of earwitnesses’ judgments of speaker identity in such cases can be presumed to be limited by the faithfulness with which speaker characteristics are conveyed over the telephone. The acoustic effects of the telephone are well-documented (e.g. Künzel 2001, Byrne and Foulkes 2004) but the auditory implications, especially those relating to voice quality, require investigation. In particular we may ask whether the perceptual distance between two voices is affected by telephone transmission. In this paper the effects of the telephone are tested by an experiment using fifteen speakers of Standard Southern British English from the Dynamic Variability in Speech (DyViS) database (Nolan et al. 2006). For each possible pairing of speakers (including same-same pairs) twenty listeners heard a short speech sample from each speaker and were asked to rate the distance between the two voices on a scale of 1 to 9. The speech samples had been recorded simultaneously in both studio and telephone quality and were heard in ‘studio only’, ‘telephone only’, and ‘mixed (telephone and studio)’ pairs. We present and interpret the results for these three conditions.

This research is part of the project ‘Voice similarity and the effect of the telephone: a study of the implications for earwitness evidence’ (VoiceSim) funded by the UK Economic and Social Research Council (RES 000-22-2582).

References


Transmission loss properties of fabrics in garments used to cover the mouth and nose

Dominic Watt¹, Philip Harrison¹, and Carmen Llamas
Department of Language & Linguistic Science, University of York, UK
JP French Associates, York, UK
dw539@york.ac.uk
pth@jpFrench.com
c1558@york.ac.uk

The acoustic transmission loss characteristics of the fabrics used for a variety of mouth/nose and face coverings - including the niqāb (full-face Muslim veil), balaclavas, and surgical masks - were compared to each other and to those of so-called 'acoustically transparent' material of the sort used for covering loudspeakers. The aim of the experiment was to obtain an estimate of the contribution made solely by the fabric masking the mouth and nose to changes in speech intelligibility that might arise when these face coverings are worn by talkers. We assess the implications of our results in a forensic phonetic context, and in the light of recent court rulings relating to the wearing of the niqāb by legal and educational professionals.
Developing the Use of Non-Traditional Methodologies in Forensic Phonetic Research: a Look at Shouted Voice Recognition in Unfamiliar Speakers

Julia Forsberg

BA (Hons) English Language with Creative Writing, Lancaster University
Forsberg.julia@gmail.com

There are many different variables which may affect the rate of successful speaker identification, both negatively and positively. These include the regional and socioeconomic accent of the speaker, together with the listener’s familiarity with these voice traits; a listener from Liverpool may for example be more able to recognise a Liverpudlian speaker. Other factors include the listening ability of the witness, if he or she has hearing problems and his or her familiarity with the voice. The pitch and frequency of the speaker may also affect the rate of recognition. Witnessing a crime has a large psychological effect on the person, and adrenaline, stress and fear (to name a few) may increase or decrease the ability to remember specific voice traits. It is therefore difficult to predict and measure the reliability of a witness. It is crucial that the procedures surrounding voice line ups (an example of these are given in the Home Office Circulation from 2003) be made as close to air-tight as can be, and that all involved parties are sufficiently educated in the field.

Keane (2006) calls for more research on the subject. From the perspective of the law, he claims that aural identification is far more dangerous in terms of the risk of misidentification than visual identification. Several cases of a phonetician warning a judge against the use of the method are cited, and Keane concludes that “Further guidance is clearly needed on aural dock identifications which, it is submitted, are no less dangerous than visual dock identifications” (Keane 2006: 258).

Künzel (1994) poses the question of ‘reliability of Speaker Identification [sic] by laypersons’ (p. 45), which is a further concern when judging what evidence should be taken into account. This is raised in this experiment as listeners are asked to specify which listener group they belong to (phonetically trained or naïve). While it is clear that trained listeners are generally more able to perform these tests, the ability of individual naïve listeners to positively identify unfamiliar voices are not inclusively less than that of trained listeners. He further discusses the consequences this type of research has on the Judiciary and concludes with suggestions for improvement of guidelines to be used in voice identification tasks. These are relevant in any Forensic Linguistics setting, due to the sometimes delicate nature of the circumstances.

An appropriate place to start is by looking at people’s actual ability to recognise voices. Most studies performed on this have involved small groups of participants; however the most efficient way of measuring how able to identify voices listeners are is to perform large scale experiments, as a complement to the small-scale experiments which have previously been the focus (Rose & Duncan 2005, Blatchford & Foulkes 2006).

These large-scale tests are performed online – this allows for a large group of listeners, and simplifies the practical problems of for example time restrictions which may otherwise arise when performing such tests. This type of methodology does of course bring its own difficulties and challenges, such as the question of trust in the listener (she may have help performing the listening activities, or she may not be who she says she is). This is however a growing part of every day life in the world, the internet is used for more and more things and it is a great opportunity to develop scientific testing and surveys and makes us able to reach a more diverse group of participants.

The website built specifically for the tests performed for this paper includes voice samples of ten speakers. The listener will before commencing the experiment enter some details for demographic
purposes, such as age, gender, regional background and possible training in linguistics/phonetics or law enforcement.

There are many aspects of voice recognition which need to be addressed in future research. One of these, the more problematic, is as mentioned above the psychological aspects. Apart from this, we will need to address: familiar vs. unfamiliar listener ability; naïve vs. phonetically trained listener ability; a listener’s background and how it affects regional and social voice trait recognition; as well as non-native speaker recognition. Another important aspect is to discuss which voice traits contribute more or less to positive voice recognition, such as accent, pitch and speech characteristics.

One of the main positives of conducting forensic research in this way is that it has the potential to reach a large number of people, and as such there is already an advantage to tests previously conducted within the field in that the sample is more diverse and representative.

References
Voice and Speech Variation under Stress

KyungWha, Kim
Forensic Science Division, Supreme Prosecutor’s Office, Korea
savoix@spo.go.kr

There is growing interest in “speech under stress” by military, and law enforcement organizations, because stress affects speech, and causes negative effect on the performance of speech technology such as speech perception, speaker recognition. Stressed speech is different from normal speech of a speaker, therefore it also affects speaker identification in forensic phonetics. This paper investigates how the voice of a speaker varies under stress and how stress affects the speaker recognition system. To instigate the stress, we use physical stress.

The research consists of 2 experiments. In experiment 1, acoustic and linguistic analysis of voice under stress-free conditions and stressed conditions were performed. In experiment 2, intra-speaker variation under stress was investigated using an Automatic Speaker Recognition (ASR) program.

For analysis on the effect of physical stress on speech production, 22 subjects (11 males, 11 females) ran up and down stairs of 6 floors for their physical stress task and read a text in 3 different conditions: before stress, immediately following stress, and 15 minutes after stress. To make sure that the subjects experienced physical change during the task, we measured physiological parameters such as blood pressure, pulse, and respiration rate.

As acoustic and linguistic analysis, we considered duration, pause, fundamental frequency (f0), rate, jitter, and speech error across the speaking conditions. To determine the influence of stressed speech on the speaker recognition system, Agnitio’s BATVOX was used.

The results of acoustic analysis indicate that duration, pause, f0, and rate were significantly different for stress and the performance of ASR program showed better recognition for non stressed speech than stressed speech.

Figure 1. mean f0 in 3 different conditions; before, right after, after stress.

Figure 2. Speaker recognition results: intra-speaker variation
Acknowledgements

The author is grateful to Dr. Hermann Künzel for his supervision and advice with this research.

References


Charlatanry and fraud – lie detectors re-visited

Anders Eriksson
Department of Linguistics, University of Gothenburg, Sweden
anders.eriksson@ling.gu.se

The present report is a follow-up to a presentation (Eriksson, 2006) given at the IAFPA annual meeting in Gothenburg, Sweden 2006. There has been considerable development in the field since then. Two independent research teams have published reports (Hollien et al. 2006, 2008; Damphousse et al. 2007) on experiments where the two most widely used so called lie or stress detectors have been tested for reliability. The results in all three studies show that the tested instruments perform at chance level.

Vendors of these instruments often claim that the lack of positive results are due to the fact that the testers have not followed the required procedures or that the experiments were carried out under conditions which are not representative of real-life situations. In all the above cited studies, the researchers therefore participated in training programs offered by the vendors to their clients. The vendors were also consulted during the studies if and when problems arose. In the study by Damphousse et al., the tested subjects were prison inmates who were interrogated about their drug use. In order to test the veracity of their answers, urinalysis tests were performed. By comparing the answers in the interrogations with the results from the urinalysis tests it was thus possible to determine with certainty the veracity of the inmates’ verbal reports. At least for this study one may thus say that the conditions matched real-life conditions perfectly. It may come as no surprise, however, that these precautions made no difference with respect to the reliability results compared to previous studies. Performance remained at chance level.

In the study by Eriksson and Lacerda (2007) no reliability tests were performed. This study is instead focused on the inner workings of the instruments in question. It is shown that the scientific claims made by the vendors of one of the two types, the so called Voice Stress Analyzer, are completely unfounded and in some cases even invented. The other type, the so called Layered Voice Analyzer, was tested by simulating the program in Mathematica. It was shown that the output has no particular connection with the human voice or what is said. The program uses statistics based on technical artefacts that arise in the digitization process as in-data. The results of these calculations are given fanciful names like “Untruthfulness”, “Low stress”, “Normal excitement”, but the assignment of labels to the statistical output is completely unmotivated.

Both types of instruments will be presented in some detail together with some suggestions on what we may do to counter this type of charlatanry in our field.

References
Due to the asylum policy in our country, which offered a temporary asylum status to anyone fleeing from Burundi in a certain period (1996-2006), many asylum seekers come to the Netherlands and claim to be from Burundi. In Burundi, Kirundi is the national language, yet there are quarters in cities like Bujumbura (the capital city) in which Swahili is the main language and the mother tongue of many speakers. Quite a number of asylum seekers claiming to be from Burundi claim to come from such a neighbourhood, and speak only Swahili. For these cases, it is essential to know: (1) to what extent people from these Swahili speaking quarters may be expected to speak some Kirundi beside Swahili, and (2) how the Swahili which is spoken in these quarters differs from other kinds of Swahili.

As it turns out, opinions differ strongly on these matters. In a large number of cases, linguists specialized in Swahili disagree with native speakers who have been trained as language analysts on what constitutes evidence for a Burundi origin. Features which are claimed by one party to be typical of Burundian Swahili are said to occur in other varieties of Swahili by the other party, and features which are claimed to point to an origin elsewhere are said to occur in Burundian Swahili as well by the other party. Unfortunately, there are no published sources about this variety of Swahili which can settle these arguments.

In order to solve this problem, our office felt the need to make recordings of native Swahili from Burundi for future publication. Last March, I went to the Swahili-speaking quarter of Buyenzi in Bujumbura and made various recordings of native Swahili. In my presentation, I will present my preliminary findings regarding phonological, morphological and lexical features of this variety of Swahili and contrast these with the features that both linguists and trained native speakers have described as typical features of Burundian Swahili in individual cases.
Language Analysis to Determine the Origin of Asylum Seekers (LADO): The 'Guidelines' in phonetic perspective

Helen Fraser
School of Behavioural Cognitive and Social Sciences, University of New England, Australia
hfraser@une.edu.au

Carol Pedersen
University of Queensland, Australia
uqcpeder@uq.edu.au

The fact that language, as well as conveying linguistic meaning, also encodes information about a speaker's regional and social identity, has been successfully used in forensic contexts for some time (Labov, 1985). In principle, such information provides a useful tool for immigration departments seeking to determine the national origin or community membership of asylum seekers presenting without papers. However, cases have been documented (Eades et al., 2003) in which language analysis used for this purpose has not conformed to acceptable standards of linguistic rigour. A set of Guidelines (LNOG, 2004) has been published, and endorsed by many linguistic organisations (McNamara, 2005, Patricks, 2008), to aid governments using language analysis for the determination of origin (LADO).

However, these Guidelines were formulated mainly on sociolinguistic considerations. Since a major indicator of origin, especially in cases where a speaker may be attempting to assume an inauthentic identity (Cambier-Langeveld, 2007) is accent, phonetic issues are also relevant. Phonetic science has been researching listeners' ability to use subtle differences of accent to identify speakers' regional and social background for many years, developing a body of knowledge about:

- what phonetic cues affect identification
- what social and personal factors affect a speaker's accent
- what social and personal factors affect a listener's ability to identify an accent
- what factors cause a speaker's accent to change in the long or short term
- the extent to which people can fake an accent other than their authentic accent
- the extent to which speakers are consciously aware of which features affect accent identification

To date, however, the question, crucial in forensic contexts, of how to judge the accuracy of accent identification under various conditions has received relatively little attention. The present paper presents background on the contexts in which LADO is undertaken in immigration cases; considers points of similarity and difference between LADO and fields such as earwitness testimony and perceptual dialectology; reviews the current state of knowledge about accent identification; and considers whether and how the Guidelines need to be extended to take account of specifically phonetic aspects of language analysis.

References

In 2004 a paper with the title "Guidelines for the use of language analysis in relation to questions of national origin in refugee cases" was published in the IJSLL. For reasons which are unclear to the present author it has in some circles received the status of guidelines for practitioners in the field.

In the call for papers Lingua, the scientific unit for linguistic analyses within the Swiss Federal office for Migration writes “In 2004, a group of linguists edited a set of Guidelines designed for all practitioners in this field. Its concern is to ensure and guarantee a sufficient quality level.” and the home page of the Taalstudio, a Dutch company performing language analysis, includes the following statement “Since June 2004, guidelines have been available for the use of Language Analysis to determine origin in asylum cases.”

There would be nothing remarkable about this if it were not for the fact that the document they refer to is not about guidelines for practitioners at all. This is explicitly stated by the authors: “The following guidelines are therefore intended to assist governments in assessing the general validity of language analysis”. We are thus led to believe that the guidelines are not intended for practitioners but for government officials evaluating analyses already performed by others. But the text which follows does not live up to this promise. Much of the paper is a rather general treatise on language variation. There are also some remarks about what qualifications an analyst should have like “Language analysis must be done by qualified linguists” and “The expertise of native speakers is not the same as the expertise of linguists” and general observations like “Language analysis requires useful and reliable data”. But there are no instructions as to how these requirements and observations may be used in an evaluation process. Assuming that the reader of the paper is a government official who has received a language analysis report and seeks advice on how to evaluate the report, it is difficult to see how these guidelines could be of much use. And as has been pointed out above, they are not even intended as guidelines for practitioners. I will therefore argue that from statements by analysts that their analyses are performed in accordance with the guidelines it is not possible to draw any definitive conclusions about what methods are actually used in the analyses.

Reading the paper leaves the reader with at least two big questions. Why is this paper referred to as guidelines by practitioners? Have they been misled by the title?

References

2) http://www.taalstudio.nl/taalanalyse/index uk.html